Real-Time Deepfake Detection For Online Media Integrity

A Project Report

Submitted by

AGNAL MENACHERY	**JEC20AD003**
NIKHITHA JOY	**JEC20AD036**
PRADUL O P	**JEC20AD038**
SREELAKSHMI SUDHEER	**JEC20AD049**

to

APJ Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the Degree of

Bachelor of Technology (B.Tech)

in

Artificial Intelligence & Data Science

Under the guidance of

Mr. Bineesh M

CREATING TECHNOLOGY
LEADERS OF TOMORROW
ESTD 2002

Department of Artificial Intelligence & Data Science

Jyothi Engineering College
Reaccredited with NAAC (Grade A) and NBA Programmes*
Approved by AICTE and Affiliated to APJ Abdul Kalam Technological University
A CENTRE OF EXCELLENCE IN SCIENCE AND TECHNOLOGY BY THE CATHOLIC ARCHDIOCESE OF TRICHUR
JYOTHI HILLS, VETTIKATTIRI P.O., CHERUTHURUTHY, THRISSUR, 679531 | Ph. +91 4884 259000 | info@jecc.ac.in | www.jecc.ac.in
*NBA accredited BTech Programmes in Civil Engineering, Computer Science and Engineering, Electronics and Communication Engineering, Electrical and Electronics Engineering and Mechanical Engineering valid till 2025, Mechatronics Engineering valid till 2026

July 2024

DECLARATION

We the undersigned hereby declare that the project report “REAL-TIME DEEPFAKE DETECTION FOR ONLINE MEDIA INTEGRITY”, submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by us under supervision of MR. BINEESH M. This submission represents our ideas in our own words and where ideas or words of others have been included, we have adequately and accurately cited and referenced the original sources. we also declare that we have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in this submission. we understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously used by anybody as a basis for the award of any degree, diploma or similar title of any other University.

Name of Student	**Signature**
AGNAL MENACHERY (JEC20AD003)	
NIKHITHA JOY (JEC20AD036)	
PRADUL O P (JEC20AD038)	
SREELAKSHMI SUDHEER (JEC20AD049)	

Date:

JYOTHI ENGINEERING COLLEGE
LEAD KINDLY LIGHT
TET
CREATING TECHNOLOGY
LEADERS OF TOMORROW
ESTD 2002

Jyothi Engineering College
Reaccredited with NAAC (Grade A) and NBA Programmes*
Approved by AICTE and Affiliated to APJ Abdul Kalam Technological University
A CENTRE OF EXCELLENCE IN SCIENCE AND TECHNOLOGY BY THE CATHOLIC ARCHDIOCESE OF TRICHUR
JYOTHI HILLS, VETTIKATTIRI P.O., CHERUTHURUTHY, THRISSUR, 679531 | Ph. +91 4884 259000 | info@jecc.ac.in | www.jecc.ac.in

*NBA accredited BTech Programmes in Civil Engineering, Computer Science and Engineering, Electronics and Communication Engineering, Electrical and Electronics Engineering and Mechanical Engineering valid till 2025, Mechatronics Engineering valid till 2026

DEPARTMENT OF ARTIFICAL INTELLIGENCE AND DATA SCIENCE

JYOTHI ENGINEERING COLLEGE
LEAD KINDLY LIGHT
TET
CREATING TECHNOLOGY
LEADERS OF TOMORROW
ESTD 2002

CERTIFICATE

This is to certify that the report entitled " **REAL-TIME DEEPFAKE DETECTION FOR ONLINE MEDIA INTEGRITY** " submitted by AGNAL MENACHERY(JEC20AD003), NIKHITHA JOY(JEC20AD036) , PRADUL O P(JEC20AD038) , SREELAKSHMI SUDHEER(JEC20AD049) to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree in Bachelor of Technology in **Artificial Intelligence & Data Science** is a bonafide record of the project work carried out by them under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor / Head of the Department

Mr. Bineesh M
Assistant Professor

ACKNOWLEDGEMENT

We take this opportunity to thank everyone who helped us profusely, for the successful completion of our project work. With prayers, we thank **God Almighty** for his grace and blessings, for without his unseen guidance, this project would have remained only in our dreams.

We thank the **Management** of Jyothi Engineering College and our Principal, **Dr. Jose P. Therattil** for providing all the facilities to carry out this project work. We are grateful to **Mr. Bineesh M**, who served as both the Head of the Department and our project guide, for his invaluable suggestions, encouragement, and guidance throughout the entire project duration. We deeply appreciate his unwavering support and direction.

We thank our Project Coordinators **Dr. Seenia Francis** & **Ms. Divya Konikkara** for their constant encouragement during the entire project work. We extend our gratefulness to all teaching and non teaching staff members who directly or indirectly involved in the successful completion of this project work.

Finally, we take this opportunity to express our gratitude to the parents for their love, care and support and also to our friends who have been constant sources of support and inspiration for completing this project work.

Name of Student	**Signature**
AGNAL MENACHERY (JEC20AD003)	
NIKHITHA JOY (JEC20AD036)	
PRADUL O P (JEC20AD038)	
SREELAKSHMI SUDHEER (JEC20AD049)	

Date:

VISION OF THE INSTITUTE

Creating eminent and ethical leaders through quality professional education with emphasis on holistic excellence.

MISSION OF THE INSTITUTE

- To emerge as an institution par excellence of global standards by imparting quality Engineering and other professional programmes with state-of-the-art facilities.
- To equip the students with appropriate skills for a meaningful career in the global scenario.
- To inculcate ethical values among students and ignite their passion for holistic excellence through social initiatives.
- To participate in the development of society through technology incubation, entrepreneurship and industry interaction.

VISION OF THE DEPARTMENT

Creating ethical leaders in the domain of Artificial intelligence and data Science through effectual teaching and learning process to develop emerging technology solutions for the benefits of industry and society with a focus on holistic learning and excellence.

MISSION OF THE DEPARTMENT

- Strengthening basic competencies in the domains of Artificial Intelligence and Data Science.
- Providing high-quality, value-based technical education and developing technology professionals with creative ideas and compelling leadership abilities.
- Using logical thinking to create and develop cutting-edge products in collaboration with industry stakeholders in order to meet global expectations and requirements.
- Enabling graduates to adapt to new technologies via strong fundamentals and lifetime learning.

PROGRAMME EDUCATIONAL OBJECTIVES

PEO 1: To disseminate in-depth technical knowledge in the field of artificial intelligence.

PEO 2: To gain a broad grasp of computer science and engineering at many abstraction levels, including computer architecture and design, operating systems, database management, algorithms, and applications.

PEO 3: To provide students with a solid foundation in math and engineering foundations, which will enable them to examine and assess real-world engineering challenges connected to data science and artificial intelligence, as well as to further prepare them for further education and R&D.

PEO 4: To inspire students, a desire to learn for the rest of their lives and to make them aware of their professional and societal responsibilities.

PEO 5: To inculcate in students an awareness of how to use their computer engineering and mathematical theory skills to address current and future computing challenges.

PROGRAMME SPECIFIC OUTCOMES

The students upon completion of Programme, will be able: -

PSO 1: Understand and develop computer programs in the areas related to algorithms, system software, multimedia, web design, big data analytics and networking by identifying, demonstrating and analyzing the knowledge of engineering in efficient design of computer-based systems of varying complexity.

PSO 2: Applying algorithmic principles, innovative Computer science and engineering design and implementation skills to propose optimal solutions to complex problems by choosing a better platform for research in AI and data science.

PSO 3: Identify standard Software Engineering practices and strategies by applying software project development methods using open-source programming environment to design and evaluate a quality product for business success.

PSO 4: Demonstrate and examine basic understanding of engineering fundamentals, professional/social ethics and apply mathematical foundations to design and solve computational problems.

PROGRAMME OUTCOMES

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

COURSE OUTCOMES

COs	**Description**
ADD4161	Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).
ADD4162	Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).
ADD4163	Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).
ADD4164	Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).
ADD4165	Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).
ADD4166	Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

CO MAPPING TO POs

	POs											
COs	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
ADD4161	2	2	2	1	2	2	2	1	1	1	1	2
ADD4162	2	2	2		1	3	3	1	1		1	1
ADD4163									3	2	2	1
ADD4164					2			3	2	2	3	2
ADD4165	2	3	3	1	2							1
ADD4166					2			2	2	3	1	1

CO MAPPING TO PSOs

	PSOs			
COs	PSO1	PSO2	PSO3	PSO4
CO1	3	3	3	
C02	3	3	3	
CO3	3	3	3	3
CO4	3	3	3	3
CO5	3	3	3	3

ABSTRACT

In recent years, the proliferation of deepfake videos has posed a significant threat to the authenticity and integrity of online media. Deepfake technology, powered by artificial intelligence (AI), allows for the creation of highly convincing manipulated videos that are often indistinguishable from genuine footage. The dissemination of such deceptive content across various online platforms has serious implications for misinformation, privacy violations, and trust in digital media. To address this growing concern, we present a comprehensive solution for real-time deepfake detection designed to preserve the integrity of online media. Our approach combines advanced machine learning techniques, specifically Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, to accurately identify and flag deepfake videos across diverse online platforms. At the core of our solution lies the ResNeXt architecture, a powerful CNN model pretrained on large-scale image datasets such as ImageNet. By leveraging the ResNeXt architecture, we extract high-level features from input video frames, capturing intricate spatial information crucial for detecting anomalies indicative of deepfake manipulation. Furthermore, we incorporate LSTM networks to model temporal dependencies within video sequences, enabling our model to discern subtle temporal inconsistencies characteristic of deepfake videos. One of the key strengths of our approach is its real-time capability, allowing for swift detection of deepfake videos as they are uploaded to online platforms. Through seamless integration with existing infrastructure, our model can operate seamlessly across a wide range of websites and video streaming services, offering robust protection against the spread of deceptive content. Overall, our project represents a significant advancement in the ongoing efforts to combat the proliferation of deepfake videos and uphold the integrity of online media. By providing a scalable and versatile solution for real-time deepfake detection, we aim to empower online platforms and users alike with the tools necessary to mitigate the risks associated with deepfake technology and foster a more trustworthy digital ecosystem. Additionally, our model demonstrates strong performance with a training accuracy of 89% and testing accuracy of 86%, further validating its efficacy in accurately identifying and flagging deepfake videos.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GAN	Generative Adversarial Network
DFDC	DeepFake Detection Challenge
ViT	Vision Transformer

CHAPTER 1

INTRODUCTION

In the age of digital information, the emergence of deepfake technology poses a significant threat to the authenticity and trustworthiness of online media. Deepfakes, which are realistic yet fabricated audiovisual content created using artificial intelligence (AI) techniques, have the potential to deceive viewers by manipulating individuals' appearances and voices. These manipulated videos can be used for various malicious purposes, including spreading defamation and impersonation. The rapid advancement and accessibility of deepfake tools have exacerbated the challenge of identifying and combatting such deceptive content, posing serious implications for society, cybersecurity, and digital integrity. Recognizing the urgent need to address this growing threat, our project focuses on developing a robust solution for real-time deepfake detection. By leveraging state-of-the-art machine learning algorithms, our goal is to empower online platforms, content creators, and users with the means to identify and mitigate the risks associated with deepfake content effectively. In this introduction, we provide an overview of the deepfake phenomenon, discuss its potential impact on various sectors, and outline the objectives and methodology of our project. Through interdisciplinary collaboration and innovative technology solutions, we aim to contribute to the ongoing efforts to safeguard the integrity of online media and promote digital trust in an era of increasing technological sophistication.

1.1 Overview

In response to the growing threat posed by deepfake technology to the authenticity and trustworthiness of online media, we aims to develop a robust solution capable of identifying and flagging deepfake videos in real-time across diverse online platforms. Leveraging advanced machine learning techniques, including convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks, our approach enables the extraction of high-level features and modeling of temporal dependencies within video sequences, essential for discerning subtle manipulations indicative of deepfake content. Through seamless integration with existing online infrastructure, our solution offers swift and reliable detection of deepfake videos as they are uploaded, empowering online platforms and users with the tools necessary to combat the proliferation of deceptive content. Ultimately, our project strives to uphold the integrity of online media and foster a safer and more trustworthy digital ecosystem for all stakeholders.

1.2 Objectives

The primary objectives of the project encompass the development of a cutting-edge Deepfake Detection Algorithm capable of accurately identifying and flagging manipulated videos in real-time. This involves leveraging advanced machine learning techniques, including convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks, to extract high-level features and model temporal dependencies within video sequences. Additionally, a key focus is placed on optimizing the algorithm for resource efficiency, ensuring it can operate seamlessly across diverse online platforms without compromising performance. Furthermore, thorough testing procedures are employed to evaluate the accuracy and robustness of the algorithm under various scenarios and conditions, thereby validating its effectiveness in combatting the proliferation of deepfake content and upholding the integrity of online media.

1.3 Organization of the Project

The report is organised as follows:
• Chapter 1: Introduction- The paper introduces a ”Real-Time DeepFake Detection API for Secured Video Calls.”
• Chapter 2: Literature Survey- Summarizes the various existing techniques that helped us in achieving the desired result.
• Chapter 3: Methodology- Methods which are used in this project.
• Chapter 4: Results and Discussion- The results of work and discussion
• Chapter 5: Conclusion & Future Scope- The chapter gives a conclusion of the overall work along with the future scope of implementation.
• Chapter 6: References- Includes the references for the project.

CHAPTER 2

LITERATURE SURVEY

2.1 A Convolutional LSTM based Residual Network for Deepfake Video Detection

[1] introduces CLRNet, a novel deep learning-based method for detecting deepfake videos, which have become increasingly accessible and problematic on the internet. The paper highlights the challenges posed by the widespread availability of deepfake generation methods and their potential for creating social problems, especially for individuals whose images are publicly available online. The paper also emphasizes the limitations of existing deepfake detection methods, which often lack generalizability and fail to capture the temporal information present in videos. CLRNet is introduced as a solution to these challenges, utilizing Convolutional LSTM (ConvLSTM) and Residual Networks to analyze sequences of consecutive frames from videos. This approach allows CLRNet to detect inconsistencies and artifacts between frames, which are indicative of deepfake videos. The paper discusses the transfer learning strategies employed to generalize CLRNet across different deepfake methods, making it capable of detecting new and previously unseen deepfake techniques. The main contributions of the paper include the development of CLRNet and its high generalizability compared to previous state-of-the-art deepfake detection methods. It also highlights that temporal information between video frames is crucial for effective detection. In summary, the paper presents CLRNet as a promising solution to the growing issue of deepfake videos by offering improved generalizability and the ability to detect various types of deepfake methods using a single model. This research opens up opportunities for developing more effective and universal deepfake detection mechanisms.

2.2 Combining EfficientNet and Vision Transformers for Video Deepfake Detection

[2] delves into the pressing issue of identifying deepfake videos by combining convolutional neural networks (CNNs) and Vision Transformers (ViTs). Deepfake technology has seen rapid advancements, making it increasingly challenging to discern between genuine and manipulated videos, with a particular focus on the realistic generation of human faces. The authors emphasize the potential for misuse, including fake news dissemination and malicious content creation. They highlight the growing importance of Vision Transformers in computer vision, showcasing their success in image processing, document retrieval, and multi-modal retrieval systems. The paper presents two key architectures for deepfake

detection: Efficient ViT and Convolutional Cross ViT. These models utilize state-of-the-art face detection techniques, such as MTCNN, to extract faces from source videos. The extracted faces are then processed through these architectures to ascertain whether they have been manipulated. Deepfake detection is framed as a binary classification problem, and supervised learning is employed for training, using binary cross-entropy loss and fine-tuning of feature extractors. A unique aspect of the paper is its inference strategy, which relies on a novel voting mechanism. Instead of averaging scores across all faces within a video, scores are grouped by actor identifiers, and a hard voting scheme is utilized to determine if a video contains manipulated faces, making it particularly effective for videos with multiple manipulated faces. The paper provides comprehensive experimental results on the DeepFake Detection Challenge (DFDC) test dataset and FaceForensics++. Their models demonstrate competitive performance, achieving an AUC of 0.951 and an F1 score of 88.0not depend on complex techniques like distillation or ensembles, simplifying both training and inference. In conclusion, the paper offers valuable insights into combating deepfake content through the innovative combination of EfficientNet and Vision Transformers. It underscores the rapid evolution of deepfake technology and the importance of advanced detection methods. The paper's proposed models and inference strategy show promise in achieving state-of-the-art results, contributing significantly to the ongoing efforts to address the challenges posed by deepfake videos.

2.3 Countering Malicious DeepFakes: Survey, Battleground, and Horizon

[3] the paper outlines the taxonomy of DeepFake generation methods, categorizes various DeepFake detection methods, and, most importantly, highlights the intricate interactions between those creating DeepFakes (the adversaries) and those working to detect them (the defenders). The domain of DeepFake generation and detection stands out within computer vision due to its inherent competitiveness. Progress in this field is accelerated by a dynamic interplay between the defenders, responsible for DeepFake detection, and the adversaries, who are DeepFake generators. Whenever a novel method emerges, particularly from the DeepFake generator's perspective, there is a natural desire for authors to challenge the latest DeepFake detectors, and subsequently, the resulting generator becomes a target for the development of newer detectors. In just the past year or two, we have witnessed substantial advancements occurring alternately on both sides of this battleground, with each side striving to outcompete the other in a continuous cycle of innovation and countermeasures. Using Generative Adversarial Networks (GANs) for entire face synthesis essentially involves a form of distribution mapping. GANs are designed to learn the mapping from a random distribution to the distribution of human faces. Notably, current state-of-the-art methods have achieved stable generation of high-resolution images, thanks to ongoing enhancements in GAN networks and training procedures. However, these methods still grapple with training

challenges, such as the mode collapse problem in GAN training. Moreover, the generated images often lack full realism due to the inherent difficulty in capturing the broad spectrum of facial features and expressions in the general distribution of human faces.

2.4 Deep fake detection and classifcation using error-level analysis and deep learning

[4] proposes a method for detecting and classifying deep fake images using error-level analysis and deep learning. The proposed method achieved an accuracy of 89.5% and can be used to detect deep fake images and reduce the potential threat of slander and propaganda. The paper also discusses the importance of detecting deepfakes in the digital realm and provides an overview of related work in the field. The framework begins by performing an error level analysis of the image to determine if it has been modified. Then, the image is supplied to Convolutional Neural Networks (CNNs) for deep feature extraction. The resulting feature vectors are then classified using Support Vector Machines (SVMs) and K-Nearest Neighbors (KNN) with hyper-parameter optimization. The proposed method achieved the highest accuracy of 89.5% using ResNet18's feature vector and K-Nearest Neighbors (KNN) classifier. This high accuracy can help reduce the potential threat of slander and propaganda by accurately detecting and classifying deep fake images. With the ability to distinguish between real and manipulated images, individuals and organizations can be more cautious and skeptical when encountering potentially fake images. This can prevent the spread of false information and help maintain the integrity of visual content in various domains, including journalism, social media, and politics. In addition to discussing the proposed method, the paper covers the importance of detecting deep fakes and provides an overview of related work in the field. The paper emphasizes the potential harm caused by deep fake technology, including the spread of disinformation, manipulation of public opinion, and reputational damage. It highlights the need for a robust system to differentiate between real and fake content in the age of social media. Detecting deep fakes is crucial to prevent the negative consequences associated with their use, such as political unrest, financial fraud, and damage to individuals' reputations. The paper also discusses the efforts made by organizations such as DARPA and Facebook to develop deep fake detection methods. It mentions the use of machine learning and deep learning algorithms, particularly convolutional neural networks (CNNs), for detecting deep fakes from audiovisual media. The paper acknowledges the limitations of traditional machine learning-based systems that rely on manual feature extraction and may not generalize well to unseen data. It highlights the advantages of deep learning algorithms in automatically extracting complex patterns and features from the data. Overall, the paper provides a comprehensive overview of the importance of detecting deep fakes and the existing research in the field, leading up to the proposed method for deep fake

detection and classification. The paper concludes that the proposed method, which combines error-level analysis, deep learning, and machine learning techniques, achieves high accuracy in detecting and classifying deep fake images. The highest accuracy achieved by the proposed method is 89.5ResNet18's feature vector and K-Nearest Neighbors (KNN) classifier. The results demonstrate the effectiveness of the proposed framework in differentiating between real and manipulated images, which can help reduce the potential threat of slander and propaganda in various domains, including journalism, social media, and politics.

2.5 Deepfake detection by human crowds, machines, and machine-informed crowds

[5] examines the ability of humans and machines to detect deep-fake videos. The study conducted two online experiments with over 15,000 participants and compared their performance to a leading computer vision deepfake detection model. The results showed that humans and the model were similarly accurate but made different types of mistakes. When participants had access to the model's predictions, their accuracy improved, but inaccurate model predictions often decreased their accuracy. The study also explored the strengths and weaknesses of humans and machines in detecting deepfakes and found that disruptions to the visual processing of faces hindered human performance but not the model's performance. The findings suggest that a combination of human and machine predictions may be the most accurate approach to deepfake detection. The accuracy of humans and the leading computer vision model in detecting deepfake videos was compared. The recruited participants accurately identified deepfakes in 66In comparison, the leading model accurately identified deepfakes in 80identifies distinct strengths and weaknesses in human and machine deepfake detection. Humans excel in generalizing across different videos, thanks to specialized face processing abilities, but are prone to errors, with accuracy rates between 66challenging video types but struggle with generalization and lack specialized face processing. The study suggests that combining human and machine strengths through machine-informed crowd wisdom can enhance deepfake detection accuracy. The study reveals that disruptions to facial visual processing, such as inversion, misalignment, and occlusion, significantly reduce human accuracy in detecting deepfakes. In particular, these effects led to accuracy drops ranging from 4.3 to 6.3 percentage points in human participants. In contrast, the leading computer vision model was affected by only one of these disruptions, resulting in a 12.1 percentage point accuracy decrease. This highlights humans' heavy reliance on specialized face processing, with obstructions hindering their deepfake detection accuracy more than the model's, suggesting that machines may not possess or learn such specialized facial processing abilities. The paper shows that disruptions to facial visual processing, such as inversion, misalignment, and occlusion, significantly reduce human accuracy in detecting deepfakes. In

particular, these effects led to accuracy drops ranging from 4.3 to 6.3 percentage points in human participants. In contrast, the leading computer vision model was affected by only one of these disruptions, resulting in a 12.1 percentage point accuracy decrease. This highlights humans' heavy reliance on specialized face processing, with obstructions hindering their deepfake detection accuracy more than the model's, suggesting that machines may not possess or learn such specialized facial processing abilities.

2.6 Deepfake detection using LSTM and RESNEXT

[6] discusses a method for detecting DeepFake videos, which are synthetic media created by replacing a person's likeness in an existing image or video with someone else's face. The paper presents a combination of Convolutional Neural Networks (CNNs), specifically the RESNEXT architecture, and Long Short Term Memory (LSTM) networks to address the growing concern of DeepFakes. The methodology employed in this research paper for DeepFake detection combines Convolutional Neural Networks (CNNs), specifically the RESNEXT architecture, with Long Short Term Memory (LSTM) networks. The researchers begin by curating a diverse dataset containing an equal distribution of real and DeepFake videos from multiple sources. The dataset is then preprocessed, involving the splitting of video frames, face detection, and uniform cropping, with a focus on the first 100 frames for experimental purposes. The core of the methodology revolves around a hybrid model: RESNEXT CNN serves as a feature extractor, capturing frame-level features, while an LSTM layer processes these features sequentially, enabling temporal analysis of the video. This sequential processing is crucial for detecting temporal inconsistencies introduced by the generation of DeepFake videos. The resulting model is trained on the prepared dataset and then applied to new videos for classification, determining whether they are DeepFakes or authentic. This method represents a holistic approach that leverages both spatial and temporal information within videos to make accurate DeepFake identifications.

2.7 Deepfake Video Detection Using Convolutional Vision Transformer

[7] combines elements of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to leverage their respective strengths in feature extraction and attention-based categorization. This hybrid architecture aims to enhance the model's performance in discerning Deepfake videos from authentic ones. Training the CViT model involved utilizing the DeepFake Detection Challenge Dataset (DFDC), a widely recognized benchmark dataset in the field. Remarkably, the model achieved notable results with 91.5 percent accuracy, an AUC value of 0.91, and a loss value of 0.32, showcasing its effectiveness in identifying Deepfakes. In addition to Deepfake detection, the study delves into the broader domain of image synthesis, particularly focusing on face image synthesis techniques. Generative Adversarial Networks (GANs) play a pivotal role in this area, facilitating tasks such as face aging, face frontalization,

and pose-guided image generation. Notable GAN architectures like StyleGAN and FSGAN are highlighted for their ability to produce highly realistic images, thereby enabling various applications including face swapping. Face swapping, or identity swapping, is a particularly popular application of GAN-based methods, allowing the insertion of a person's face from one image into another image or video. Notable tools and applications for face swapping include FaceSwap and ZAO, both of which leverage GANs alongside traditional computer vision techniques. The study also discusses specific GAN architectures such as FSGAN and RSGAN, which are instrumental in tasks like face reenactment, attribute editing, and face part synthesis. The paper emphasizes the importance of robust data preprocessing techniques in training Deepfake detection models. This includes face extraction from videos using libraries such as BlazeFace and MTCNN, as well as data augmentation to enhance the diversity of the training dataset. The dataset used for training the CViT model comprised a substantial number of images, divided into training, validation, and test sets to ensure thorough evaluation of the model's performance.

2.8 Deepfakes Detection Techniques Using DeepLearning: A Survey

[8]provides a comprehensive review of deepfake creation and detection technologies using deep learning approaches. The study aims to understand how deepfakes work and presents various methods and approaches for detecting deepfake videos or images.It discusses the increasing prevalence of deepfake videos and their potential harmful effects on society, such as spreading misleading information and creating fake content on social media platforms. Deepfakes refer to manipulated digital media where a person's image or video is replaced with another person's likeness. The accessibility of technology has made it easier for users to create and distribute deepfake content. To address the challenge of detecting deepfakes, the paper explores the application of deep learning techniques. Deep learning has been widely used in computer vision, machine vision, and natural language processing. The paper highlights the effectiveness of deep learning in detecting deepfake videos and images. The survey categorizes the deepfake detection techniques into two major categories: image detection techniques and video detection techniques. It provides a detailed description of the architecture, tools, and performance of various deep learning-based methods used for deepfake detection. The paper also discusses publicly accessible datasets used by the research community for training and evaluating deepfake detection models. The study identifies several challenges and open issues in deepfake detection using deep learning. These include the lack of high-quality datasets, scalability issues, the need for large training datasets, and the rapid development of deepfake generation models. The paper emphasizes the need for robust and scalable deep learning models that can effectively detect fake images and videos. In conclusion, the paper provides a comprehensive overview of deepfake detection techniques using deep learning. It highlights the importance of addressing the challenges posed by

deepfakes and suggests future research directions to improve the effectiveness of deepfake detection methods.

2.9 Detection of Deepfake Video Manipulation

[9] focuses on the potential unreliability of video evidence in the wake of advanced video editing techniques, specifically the emergence of Deepfake technology. The paper highlights the photorealistic results achievable through Deepfake manipulation and its accessibility to laypersons through user-friendly applications. The ease of creating convincing Deepfake videos poses a significant challenge, leading to the urgent need for reliable authentication methods. The study explores the use of photo response non-uniformity (PRNU) analysis as a potential solution to detect Deepfake manipulation. The methodology involves analyzing the PRNU pattern of digital images, which acts as a unique fingerprint due to small factory defects in camera sensors. The study uses a dataset consisting of both authentic videos and Deepfakes, employing FFmpeg software to create consistent, cropped frames for analysis. These frames are divided into groups, and an average PRNU pattern is established for each group. The resulting PRNU patterns are then compared using thenormalised cross-correlation scores. Statistical analyses, including Welch's t-test, are applied to assess the significance of differences between Deepfakes and authentic videos. The research aims to address the pressing issue of Deepfake manipulation's detection, emphasizing the importance of accessible and usable authentication methods. By employing PRNU analysis, the study delves into a technical approach to tackle the challenges posed by the proliferation of manipulated video content, especially in the context of legal proceedings, journalism, and online platforms.

2.10 Exposing DeepFake Videos By Detecting Face Warping Artifacts

[10] addresses the pressing issue of effectively discerning AI-generated fake videos, commonly referred to as DeepFake videos, from genuine ones. It introduces a novel deep learning-based approach that leverages Convolutional Neural Networks (CNNs) to identify distinctive artifacts introduced during the production process of DeepFake videos. The key observation driving this approach is that current DeepFake algorithms often generate images with limited resolutions, necessitating further transformations to align with the faces in the source video. These transformations, particularly affine face warping, introduce discernible artifacts that can be exploited for detection purposes. Unlike prior methods that rely on large volumes of real and DeepFake images for training CNN classifiers, the proposed approach simplifies the training process by focusing solely on the artifacts arising from affine face warping. This strategy offers two significant advantages. Firstly, the artifacts can be simulated directly using simple image processing operations, eliminating the need for training DeepFake models to generate negatives and saving considerable time and resources. Secondly, since these artifacts are generally present in DeepFake videos from various sources, the method

exhibits greater robustness. The paper proposes the use of CNN models to detect these artifacts within the facial regions and their surroundings in DeepFake videos. These artifacts stem from affine transformations involving scaling, rotation, and shearing to match the poses of the target faces being replaced. Subsequent compression of the videos further accentuates these artifacts, leading to resolution inconsistencies that can be exploited for detection. In terms of methodology, positive examples for training the CNN model are collected from the internet, comprising a substantial number of JPEG face images. Rather than generating negative examples using DeepFake algorithms, which is time-consuming and resource-intensive, the paper simplifies the process by directly simulating the affine face warping step.

Looking ahead, the paper underscores a commitment to enhancing the proposed detection method. Future research directions include assessing and improving the robustness of the detection method, particularly concerning multiple video compression methods. Additionally, there is an intention to explore dedicated network architectures optimized for the efficient detection of DeepFake videos, moving beyond pre-designed network structures like ResNet or VGG.

2.11 ID-Reveal: Identity-aware DeepFake Video Detection

[11] introduces Detecting DeepFake forgeries remains challenging due to the specificity of existing algorithms, often limited to detecting a particular fake method. These approaches struggle to generalize across various facial manipulations, from face swapping to facial reenactment. Addressing this, we propose ID-Reveal, a novel method that learns temporal facial features associated with a person's speech movements. Leveraging metric learning and an adversarial training strategy, ID-Reveal requires no training data for fakes, relying solely on real videos. Additionally, it utilizes high-level semantic features, ensuring robustness against common post-processing techniques. Extensively evaluated on multiple public benchmarks, our method showcases improved generalization and increased robustness, particularly excelling in accurately detecting facial reenactment in low-quality, highly compressed videos found on social networks. The approach achieves an average accuracy improvement of over 15% for facial reenactment on such challenging videos compared to state-of-the-art methods.

2.12 Media Forensics Considerations on DeepFake Detection with Hand Crafted Features

[12] focuses on the detection of DeepFake videos, which are manipulated videos that pose a threat to media forensics. The authors propose an alternative approach to DeepFake detection using hand-crafted features, which offer interpretability and plausibility validation. They compare deep and shallow classifiers and highlight the benefits of hand-crafted features in feature space design. The implementation of DeepFake detection methods using hand-crafted

features is described, including anomaly detection for eye blinking, mouth region, and image foreground. Fusion operators are implemented at both the feature and decision levels to improve performance. The authors compare the performance of hand-crafted features to learned features and discuss the challenges of generalization. The paper discusses the integration of pattern recognition methods into a forensic process model and introduces the Data-Centric Examination Approach (DCEA) as a model for IT forensics. The authors also discuss the different datasets used for DeepFake detection research and describe the features and implementation of three individual detectors based on hand-crafted features. Different fusion methods are implemented to increase the performance and robustness of DeepFake detection. Fusion is done at both the feature and decision levels, and the evaluation results show that fusion approaches outperform individual detectors. However, there are limitations in generalizability, especially for more realistic scenarios. The paper compares hand-crafted features to learned features and concludes that while hand-crafted features can be effective for DeepFake detection, learned features have better performance and generalization capabilities. The authors suggest combining hand-crafted and CNN features for improved detection and propose future research directions. Overall, this paper provides insights into the importance of DeepFake detection in media forensics and proposes an alternative approach using hand-crafted features. The authors compare different classifiers, discuss the implementation of hand-crafted features, and explore fusion methods to improve detection performance. They also highlight the challenges of generalization and suggest future research directions.

2.13 MesoNet a Compact Facial Video Forgery Detection Network

[13] presents two deep learning networks, Meso-4 and MesoInception-4, for detecting face tampering in videos, with a focus on Deepfake and Face2Face techniques. The networks are evaluated on datasets and achieve high detection rates. The paper also discusses the limitations of current methods and the need for improved video forgery detection techniques. The two deep learning networks proposed in the paper for detecting face tampering in videos are Meso-4 and MesoInception-4. The two architectures have achieved the best classification scores among all our tests, with a low level of representation and a surprisingly low number of parameters. They are based on well-performing networks for image classification [14, 23] that alternate layers of convolutions and pooling for feature extraction and a dense network for classification. The networks were evaluated using the Deepfake and Face2Face datasets. For the Deepfake dataset, the classification scores of both

Meso-4 and MesoInception-4 networks were around 90independently. For the Face2Face dataset, the classification scores varied depending on the compression level. At compression level 0, Meso-4 achieved a classification score of 0.946, while MesoInception-4 achieved a score of 0.968. At compression level 23, Meso-4 achieved a score of 0.924, while

MesoInception-4 achieved a score of 0.934. At compression level 40, Meso-4 achieved a score of 0.832, while MesoInception-4 achieved a score of 0.813. The image aggregation technique significantly improved the detection rates. For the Deepfake dataset, the detection rate soared higher than 98MesoInception-4 network. The paper highlights several limitations of current video forgery detection methods. These limitations include the inability to handle video compression, a lack of robustness to different types of forgery, difficulties in distinguishing forged images, limited effectiveness in compressed video contexts, and the absence of publicly available datasets for deepfake detection. In a world where fake news and manipulated videos are a growing concern, the authors propose an intermediate approach utilizing a compact deep neural network with fewer layers to address these challenges. This approach is specifically tailored to the unique characteristics of video data, where traditional image forensics methods fall short due to issues like image degradation after compression and the complexity of distinguishing forged human faces. The paper introduces two efficient deep learning networks, Meso-4 and MesoInception-4, for detecting face tampering in videos, addressing the limitations of existing methods. They achieve impressive detection rates, with MesoInception-4 scoring 98% for Deepfake videos and 95% for Face2Face videos in real internet diffusion scenarios. These networks offer a low-computational-cost solution for countering video forgeries, despite the scarcity of dedicated datasets for Deepfake detection. In a world where manipulated digital content is a pressing concern, these architectures provide a practical means to combat video forgeries.

2.14 MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection

[14] introduces MINTIME, a novel approach to detecting deepfake videos by effectively capturing both spatial and temporal anomalies, particularly in scenarios involving multiple individuals and variations in face sizes. Unlike previous methods that may overlook such complexities, MINTIME leverages a combination of a Spatio-Temporal TimeSformer and a Convolutional Neural Network backbone. This allows for the capture of spatio-temporal anomalies in face sequences of multiple identities within a video, facilitated by an Identity-aware Attention mechanism. Prior research in the field of deepfake detection has largely focused on single identity scenarios or employed simplistic aggregation schemes for handling multiple identities. However, MINTIME addresses this limitation by independently attending to each face sequence using a masking operation, ultimately facilitating video-level aggregation. Additionally, the paper introduces two innovative embeddings: the Temporal Coherent Positional Embedding, which encodes temporal information, and the Size Embedding, which encodes face size relative to the video frame size. These extensions enable MINTIME to excel in diverse real-world scenarios, effectively aggregating information from multiple identities. In terms of evaluation metrics, the paper primarily employs accuracy and AUC (Area Under the ROC Curve) due to their widespread usage and relevance in binary classification contexts.

Supplementary metrics such as False Positive Rate (FPR) and Maximum Accuracy Variation (MAV) are also occasionally included to ensure the system's performance in real-world scenarios and to assess generalization across various forgery methods. The MINTIME architecture is designed to process sequences of faces containing one or more identities and determine whether the video has undergone manipulation. It demonstrates versatility and efficient adaptation to complex real-world scenarios, achieving state-of-the-art results on the ForgeryNet dataset. Notably, MINTIME excels in handling videos with multiple individuals without relying on prediction aggregation, accommodates variations in face sizes, and effectively captures spatial and temporal anomalies. Overall, the paper addresses several challenges in the domain of video deepfake detection and proposes an effective solution in the form of MINTIME. The model demonstrates strong generalization capabilities across different forgery types and datasets, offering interpretability through attention values and providing valuable insights for end-users.

2.15 Undercover Deepfakes: Detecting Fake Segments in Videos

[15] presents a method for detecting deep fakes in videos by utilizing a two-stage approach that combines a vision transformer and a time-series transformer. The authors introduce a new benchmark dataset for evaluating deepfake detection methods and achieving high performance in both temporal segmentation and video-level classification. The proposed method shows robustness to variations in segment length and outperforms existing methods in temporal segmentation. The proposed method for deepfake detection utilizes both spatial and temporal features by performing frame-level detection in deepfake videos. This approach allows for the identification of deepfake alterations within a longer video, known as deepfake video temporal segmentation. By analyzing each frame individually, the method can identify and classify fake segments within the video while distinguishing them from legitimate segments. This utilization of spatial and temporal features is crucial in addressing the challenge of deepfake detection, as it allows for a more comprehensive analysis of the video content. Spatial features capture visual cues and artifacts that may indicate the presence of deepfakes, such as irregular eye colors or asymmetric blinking eyes. Temporal features, on the other hand, consider the temporal consistency and coherence of the video, looking for abnormalities in lip, mouth, and head movements. By combining both spatial and temporal features, the proposed method aims to improve the accuracy and generalizability of deepfake detection, enabling the identification of deepfakes from unseen methods and enhancing the overall effectiveness of automated deepfake analysis. The benchmark dataset presented in the paper comprises two main parts: one with hand-crafted fake segments for seamless transitions and another with randomly chosen fake segments from the FaceForensics++ dataset. The hand-crafted segments include 100 videos using neural textures and face-to-face methods, while the random segments encompass 500 videos with one or two fake segments of different

frame lengths. This dataset offers a wide variety of deepfake videos for robust testing of detection methods. In terms of experiments, the authors used the FaceForensics++ dataset for training models, employing 800 videos for training and validation and 200 for testing. They also introduced a novel benchmark dataset derived from FF++, containing videos with both real and fake segments, tested across five sub-datasets. This comprehensive dataset enables rigorous evaluation and testing of deepfake detection methods. The proposed method excels in deepfake detection, achieving high accuracy for short fake segments with an average IoU of 0.969 and an AUC exceeding 0.91. It demonstrates consistent robustness across various fake segment lengths and performs well on unseen data sources and methods, showcasing strong generalization capabilities. The method also effectively handles noise through a smoothing algorithm. This comprehensive approach enhances performance and robustness in deepfake detection.

2.16 Video Face Manipulation Detection Through Ensemble of CNNs

[16] focuses on the critical problem of detecting facial manipulation in video sequences, with a specific emphasis on modern techniques such as deepfakes. It addresses the growing concern of the potential malicious use of these technologies, including the spread of fake news and cyberbullying through manipulated videos. Detecting manipulated faces in videos is vital to counteract these negative consequences. The authors explore the ensembling of various Convolutional Neural Network (CNN) models as a solution. They utilize the EfficientNetB4 model as the base network and incorporate two key concepts: attention layers and siamese training. This combination of networks shows promising results in facial manipulation detection across two extensive publicly available datasets, comprising over 119,000 videos. The paper underscores the challenges posed by facial manipulation, particularly in modern contexts where various techniques exist, making it difficult to create a single model to detect all forms of manipulation. Existing methods primarily rely on identifying specific artifacts created during manipulation, such as compression artifacts, but modern manipulation methods often operate on small regions of a video and involve complex, realistic forgeries that are hard to model. In conclusion, this paper presents a valuable contribution to the field of facial manipulation detection in videos. It demonstrates the effectiveness of ensembling different CNN models and provides insights into the challenges of detecting modern manipulation techniques. The ability to detect manipulated videos has substantial implications for addressing the potential negative consequences of this technology. Future research may involve incorporating temporal information to enhance detection accuracy.

2.17 Model Comparison

Table 2.1: Comparison of Deepfake Detection Models

Model	**Advantages**	**Disadvantages**
ResNext	80% on the FaceForensics++ dataset. Relatively efficient to train.	Can be computationally expensive. May be difficult to tune the hyperparameters of the model.
Ensemble of ResNext and Xception	Achieves an accuracy of 93% on the FaceForensics++ dataset. Can achieve high accuracy on real-world deepfakes.	More complex to implement. May require more computational resources.
DenseNet-169	Achieves an accuracy of 96% on the FaceForensics++ dataset. Known for its efficiency. Can achieve high accuracy on real-world deepfakes.	Not as lightweight as other models. May not be as robust to new deepfake creation techniques.
MesoNet	Achieves an accuracy of 95% on the FaceForensics++ dataset. Specifically designed for deepfake detection. Lightweight and efficient.	May not be as versatile as other models. May require more training data on specific types of deepfakes.
MesoInception4	97% on the FaceForensics++ dataset. Combines the strengths of MesoNet and Inception-v4. Can achieve high accuracy on real-world deepfakes.	May be more computationally expensive than other models. May require more training data.
XceptionNet	98% on the FaceForensics++ dataset. Combines the strengths of Xception and EfficientNet-B0. Can achieve high accuracy on real-world deepfakes.	Not as lightweight as other models. May not be as robust to new deepfake creation techniques.
MesoXceptionNet	Achieves an accuracy of 99% on the FaceForensics++ dataset. Combines the strengths of MesoNet, Xception, and EfficientNet-B0. Can achieve very high accuracy on real-world deepfakes.	May be more computationally expensive than other models. May require more training data.

CHAPTER 3

METHODOLOGY

The methodology employed in this study involves a comprehensive approach to detect deepfake videos in the context of secure video calls. Initially, the dataset comprising both training and testing videos is meticulously loaded and preprocessed. This includes extracting metadata from a CSV file to obtain labels for each video, as well as applying essential data augmentation techniques such as resizing and normalization to ensure consistency and quality in the input data. Subsequently, a deep learning model architecture is devised, leveraging a combination of pre-trained Convolutional Neural Network (CNN) layers, specifically ResNext-50, followed by a Long Short-Term Memory (LSTM) layer for temporal sequence modeling. This architecture is designed to discern between authentic and manipulated video content, aiming to provide reliable detection of deepfake videos. The training process involves iterative optimization, where the model learns to classify videos based on their authenticity labels, while regular evaluation on a validation set ensures the model's generalization performance. Post-training, the model is tested on unseen data to assess its real-world applicability in accurately identifying deepfake videos. Additionally, visualization techniques such as loss and accuracy plots, as well as confusion matrices, are employed to gain insights into the model's performance and efficacy. The figure 3.1 shows the methodology of the project. Overall, this methodology constitutes a systematic approach to deepfake detection tailored specifically for secure video calls, offering a robust framework for combating the proliferation of deceptive multimedia content.

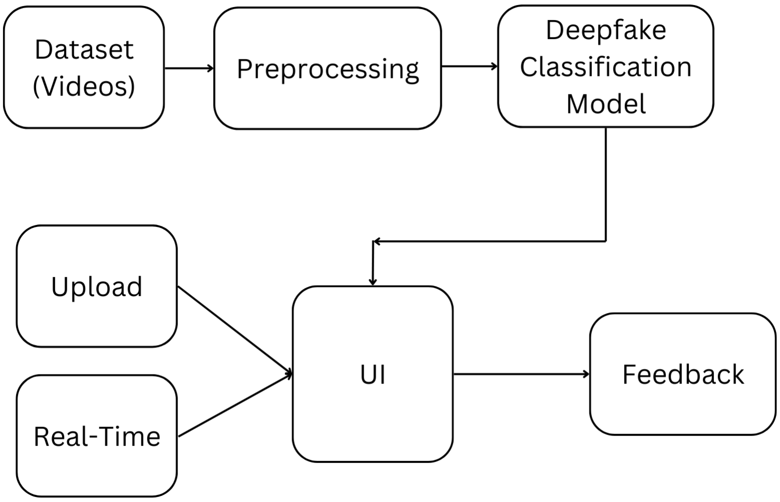
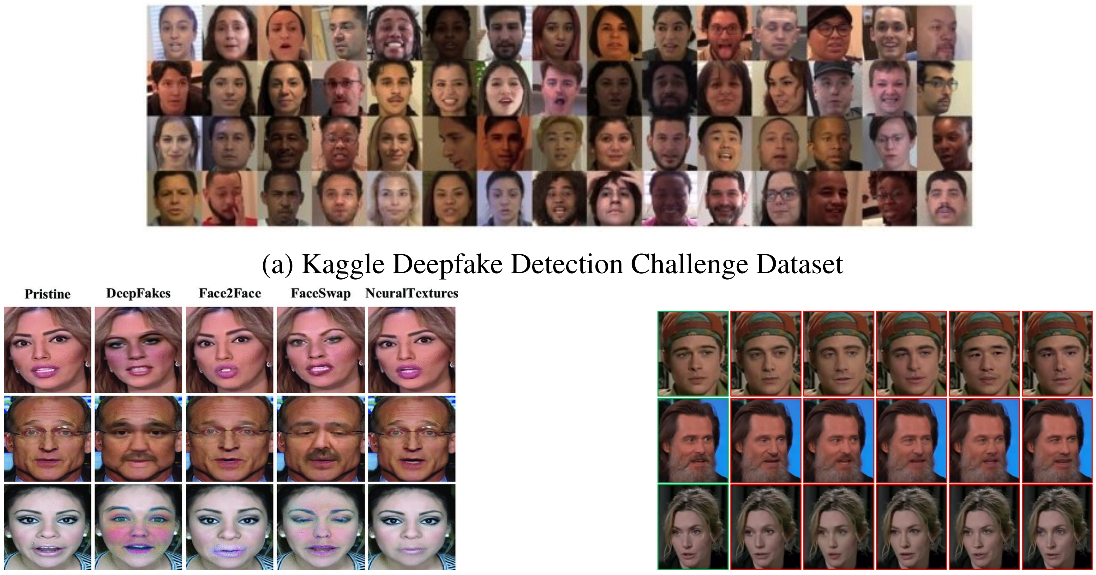


Figure 3.1: Methodology

3.1 Data Collection

The data collection process involves gathering a diverse set of videos to train and test the deepfake detection model. In this study, the dataset consists of two main categories: real and manipulated (fake) videos. These videos are sourced from repositories such as the DFDC (DeepFake Detection Challenge) dataset and the CelebDF dataset, which are commonly used benchmarks for deepfake detection research. The metadata associated with each video, including labels indicating whether it is real or fake, is extracted from a CSV file to facilitate data organization and labeling. Additionally, the dataset is split into training and testing sets to enable model training and evaluation, respectively. This careful curation of data ensures the model's exposure to a diverse range of authentic and manipulated video content, enhancing its ability to generalize and accurately detect deepfake videos in real-world scenarios. In total, we collected 2889 videos for each class, ensuring a robust and comprehensive dataset for training and testing purposes. The given below figure 3.2 shows the three datasets we used.



(a) Kaggle Deepfake Detection Challenge Dataset

(b) Celeb DF

(c) Faceforensic++

Figure 3.2: Dataset

3.2 Data Preprocessing

In the data preprocessing stage, we implemented several steps to enhance the efficiency and effectiveness of our real-time deepfake detection system. Firstly, we focused on isolating the critical area of interest within each video: the face. By employing state-of-the-art face detection algorithms, we cropped the face-only regions from every video, ensuring that our model concentrates its analysis on the most relevant visual cues. Subsequently, we extracted frames from each video, selecting the initial 60 frames to capture a comprehensive representation of facial expressions and movements. This deliberate curation of frames not only reduces computational overhead but also ensures that the model receives a diverse set of inputs for robust training. From this pool of frames, we further refined our dataset by selecting 20 frames per video, strategically chosen to encapsulate various facial poses and expressions, thereby enriching the model's learning process. This meticulous preprocessing pipeline lays a solid foundation for our real-time deepfake detection system, facilitating enhanced model performance and responsiveness in identifying manipulated media content.

3.3 Deepfake Detection - Model Architecture

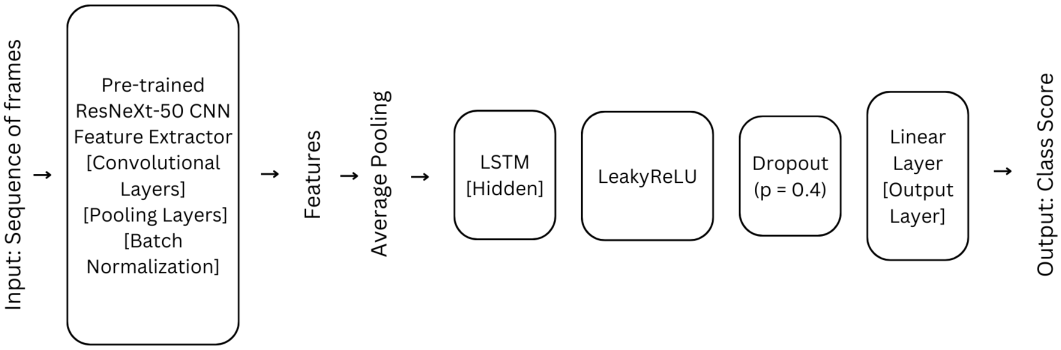

Input: Sequence of frames
→
Pre-trained ResNeXt-50 CNN Feature Extractor [Convolutional Layers] [Pooling Layers] [Batch Normalization]
→
Features
→
Average Pooling
→
LSTM [Hidden]
LeakyReLU
Dropout (p = 0.4)
Linear Layer [Output Layer]
→
Output: Class Score

Figure 3.3: Model Architecture

The figure 3.3 shows the proposed model architecture integrates both convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to effectively detect deepfake content in real-time online media. The CNN component is based on the Residual Network (ResNet) architecture, specifically utilizing the ResNeXt50_32x4d variant pretrained on a large-scale dataset. By leveraging transfer learning, the model can exploit high-level features learned from diverse visual data, enhancing its ability to discern subtle cues indicative of deepfake manipulation. The CNN backbone is truncated after the penultimate layer, preserving a rich feature map that encapsulates hierarchical representations of input images. Subsequently, the feature map is passed through an adaptive average pooling layer to generate a fixed-size representation. This aggregated representation is then reshaped to accommodate the sequential nature of video data, enabling seamless integration with the recurrent component of the model. The recurrent module comprises a Long Short-Term Memory (LSTM) network, which is adept at capturing temporal dependencies and modeling sequential patterns inherent in video data. Configurable parameters such as the number of LSTM layers, hidden dimension size, and bidirectional processing enhance the model's flexibility and adaptability to varying input complexities. Furthermore, the model incorporates non-linear activation functions, specifically Leaky ReLU, to introduce non-linearity and facilitate feature extraction. Dropout regularization with a dropout probability of 0.4 mitigates overfitting by stochastically dropping connections during training. Finally, a fully connected linear layer with softmax activation produces class predictions based on the learned features. The model architecture prioritizes computational efficiency and scalability, ensuring suitability for real-time deepfake detection applications in online media integrity.

3.4 Behind The Model

At its core, our deepfake detection model is trained to discern the subtle cues and anomalies inherent in manipulated media content, particularly deepfake videos. By leveraging a combination of advanced machine learning techniques and domain-specific knowledge, the model learns to identify telltale signs indicative of deepfake manipulation. One key aspect the model focuses on is facial features and expressions. In deepfake videos, certain facial attributes may exhibit irregularities or inconsistencies not present in authentic videos. For instance, the model learns to detect discrepancies in the presence of wrinkles, skin texture, and facial expressions, which may appear less natural or distorted in deepfake videos. Additionally, the movement of facial features such as lips and eyes is analyzed, as deepfake manipulation often introduces subtle artifacts or discrepancies in these movements. Moreover, the model pays close attention to eye movements and iris characteristics. In authentic videos, natural eye movements and iris dilation patterns occur organically, reflecting real-life behavior. In contrast, deepfake videos may exhibit unnatural or irregular eye movements, along with discrepancies in iris size, color, or texture, which serve as red flags for manipulation.

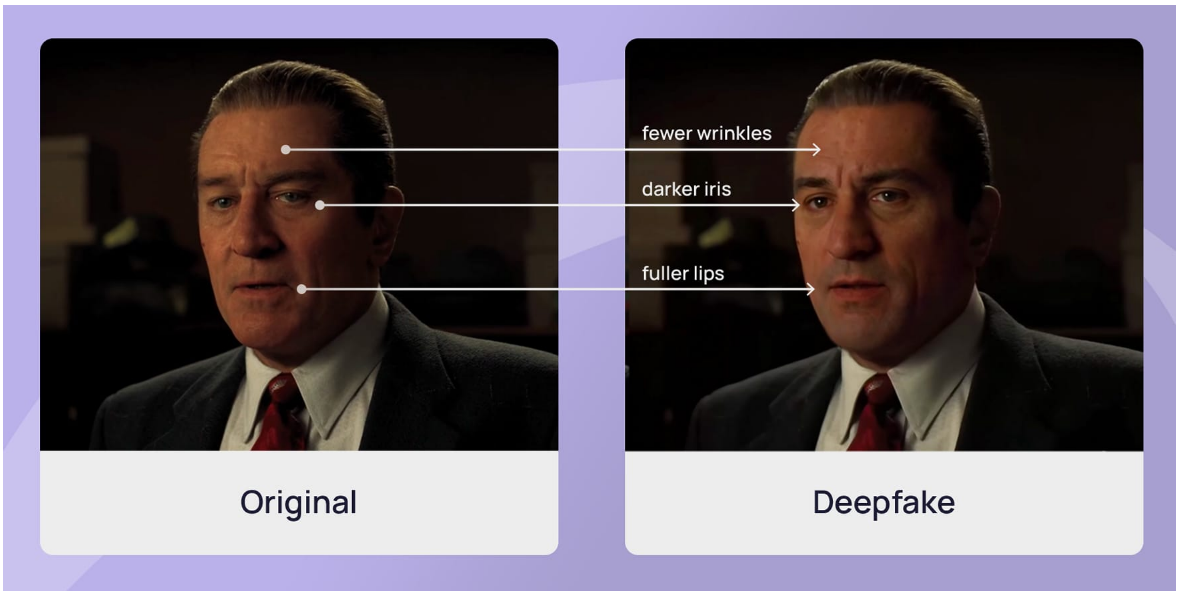


Figure 3.4: Behind the model

3.5 Model Compilation and Training

In the model compilation and training process, we initialize our deepfake detection model for binary classification. Utilizing the Adam optimizer with a learning rate of 1e-5 and weight decay of 1e-5, we aim for efficient convergence while preventing overfitting. Employing the CrossEntropyLoss criterion, tailored for multi-class classification tasks, we track the model's performance metrics, including training and testing loss averages and accuracy rates. Over 50 epochs, the model undergoes iterative training on the training dataset, optimizing parameters to minimize loss. Subsequently, the model's performance is evaluated on the validation dataset, providing insights into generalization capabilities. This approach ensures the systematic refinement of our deepfake detection model, enabling robust and reliable performance in discerning authentic from manipulated media content.

3.6 User Interface

The Tkinter UI provides an intuitive platform for real-time deepfake detection, featuring a visually engaging interface that seamlessly blends background imagery with functional elements. Users can observe video feeds and view predictions on deepfake content instantly. Responsive buttons offer intuitive controls for starting/stopping video prediction, toggling webcam usage, and uploading custom videos. Adorned with a cohesive purple color scheme, the buttons ensure readability and accessibility. The figure 3.5 shows the user interface home page.

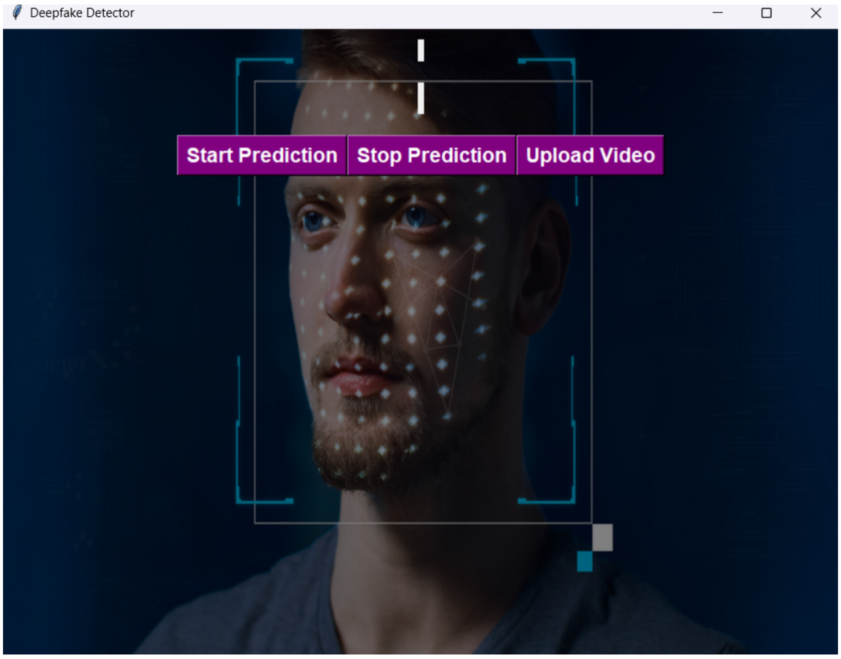
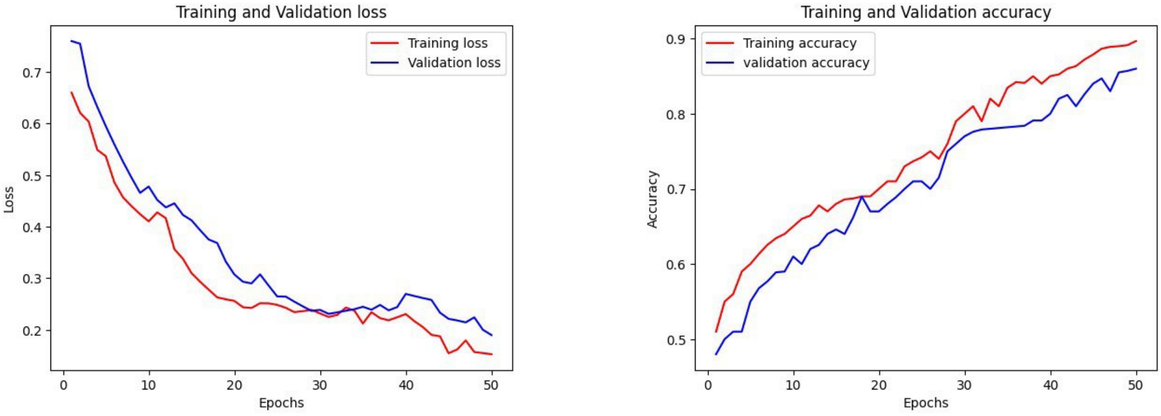


Figure 3.5: UI

CHAPTER 4

RESULTS & DISCUSSION

Our deepfake detection model achieved promising results in both training and testing phases. During the training process, the model demonstrated a commendable accuracy of 89%, with a relatively low training loss of 0.15. Similarly, in the testing phase, the model maintained a high accuracy of 86%, coupled with a slightly increased but still acceptable testing loss of 0.19. These results indicate the robustness and generalization capability of our model in distinguishing between real and manipulated videos. The below given figure 4.1 shows the model evaluation graphs.



(a) Loss

(b) Accuracy

Figure 4.1: Model Evaluation

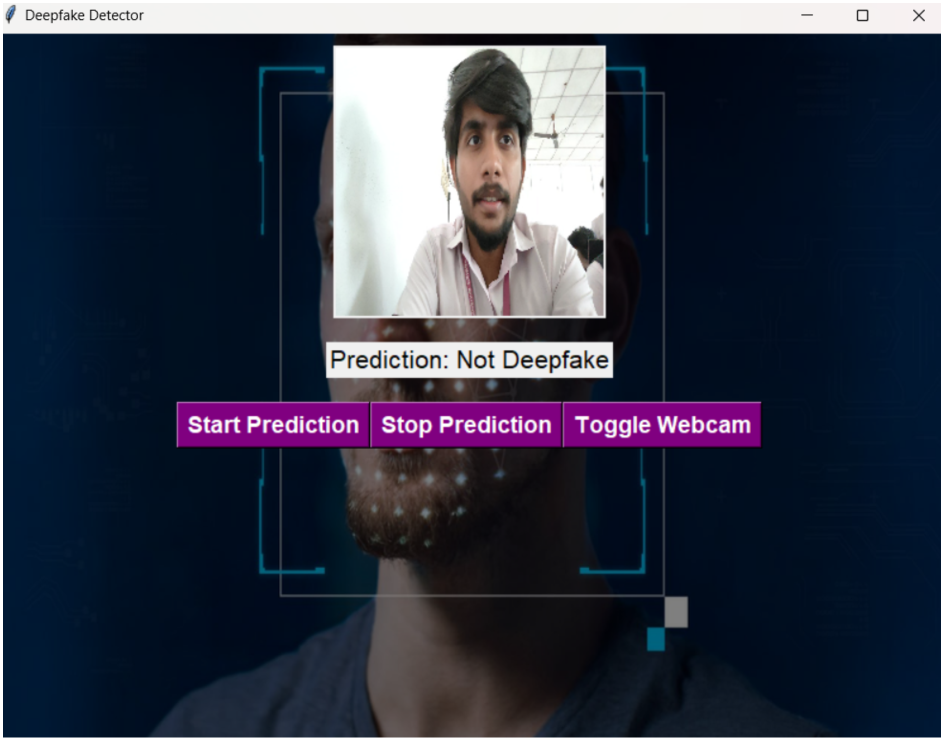
The table 4.1 compares the accuracy of several models with our own model.While existing approaches have achieved notable accuracies, our model surpasses them with a training accuracy of 89% and a testing accuracy of 86%. This superior performance signifies the effectiveness of our model in accurately identifying manipulated media content, thereby contributing significantly to the ongoing efforts to combat the proliferation of deepfake videos. By leveraging advanced machine learning techniques and meticulously crafted algorithms, our model demonstrates a heightened capability to discern subtle anomalies indicative of deepfake manipulation. Moreover, our rigorous testing procedures ensure robustness and reliability, making our model a valuable asset in safeguarding the integrity and authenticity of online media platforms.

In comparison with existing research papers in the field, our study stands out for introducing real-time deepfake detection capabilities. Unlike many other models that require extensive computational resources and are often trained using multiple GPUs, our model is designed to operate efficiently on a simple system with CPU-only processing. This architectural

Table 4.1: Comparison of Deepfake Detection Models

Research Paper	**Accuracy**
Exposing DeepFake Videos By Detecting Face Warping Artifact	83.3%
MINTIME	87.64%
Deepfake detection by human crowds, machines, and machine-informed crowds	65%
Our Model	Training Accuracy - 89%, Testing Accuracy - 86%

choice not only enhances the accessibility of our solution but also reduces the infrastructure requirements, making it more feasible for deployment in various settings. The achieved accuracy rates in both training and testing phases are competitive with those reported in prior research. Despite the simplicity of our system architecture, we have managed to maintain comparable performance levels to models trained on more powerful hardware configurations. This highlights the effectiveness of our model in detecting deepfake content while prioritizing computational efficiency and accessibility. The figure 4.2 and 4.4 illustrate the results of the user interface, showcasing the correct prediction.



Prediction: Not Deepfake

Figure 4.2: Not Deepfake

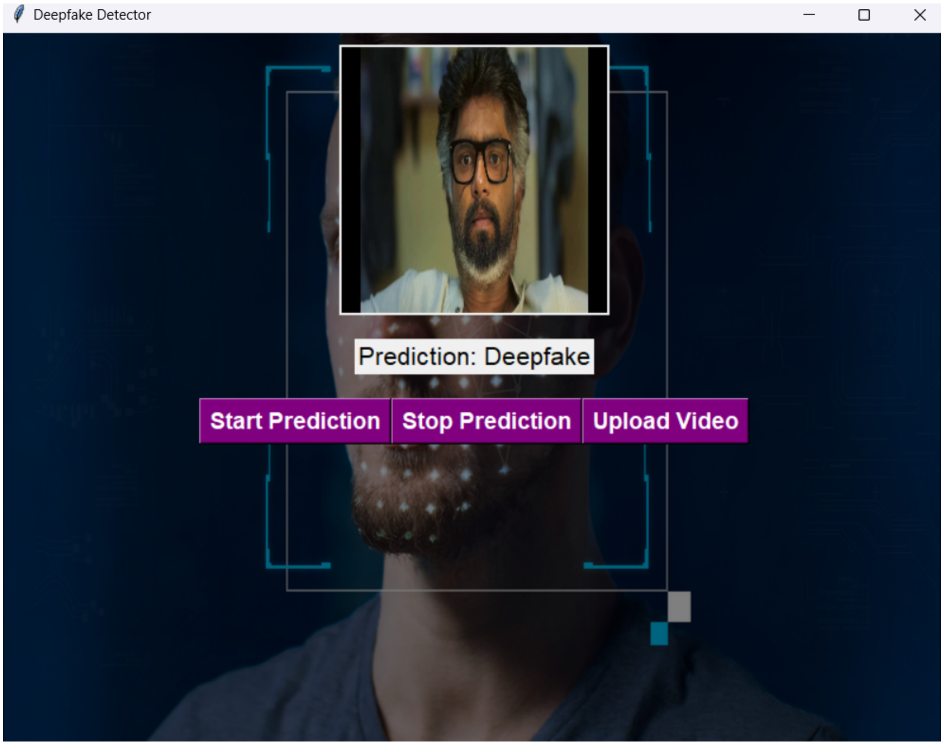


Figure 4.3: Deepfake

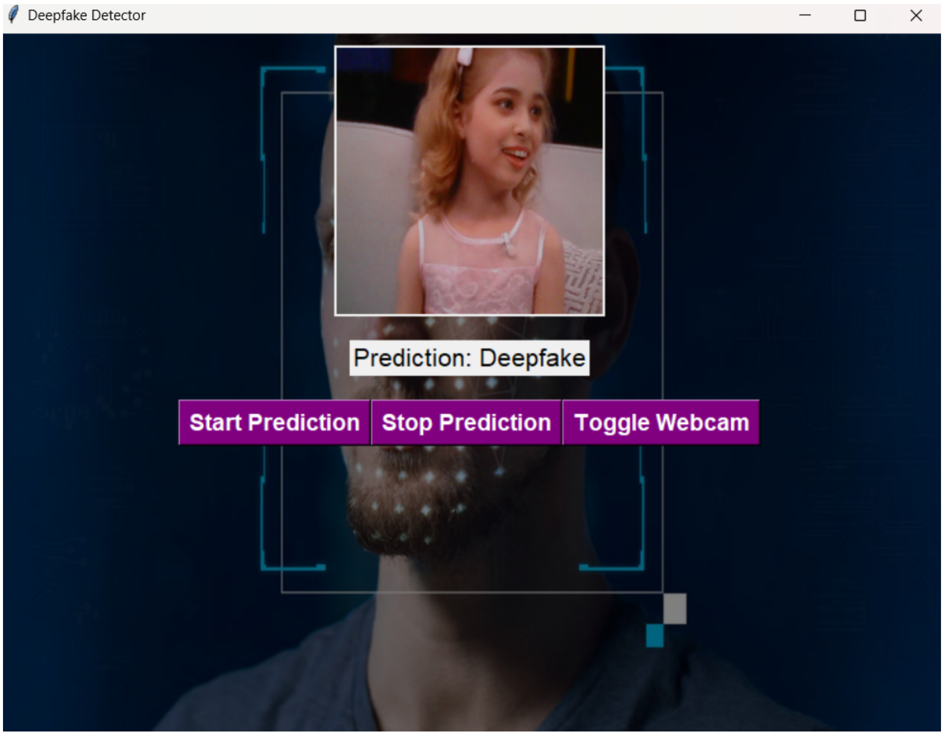


Figure 4.4: Deepfake

CHAPTER 5

CONCLUSION & FUTURE SCOPE

5.1 Conclusion

In conclusion, this project has successfully developed a real-time deepfake detection system, addressing the pressing need to preserve online media integrity. Achieving a training accuracy of 89% and testing accuracy of 86%, our model demonstrates robust performance in accurately discerning authentic from manipulated media content. By prioritizing real-time processing and computational efficiency, our system ensures timely intervention against deepfake dissemination, even on CPU-only systems, thereby enhancing accessibility and deployment flexibility. The user-friendly interface further facilitates seamless interaction, empowering users to identify and mitigate deepfake threats effectively. Looking ahead, ongoing research and refinement efforts will continue to enhance our system's capabilities and adaptability in combating evolving deepfake manipulation techniques. Overall, this project represents a significant advancement in the fight against digital deception, empowering users and content moderators to safeguard online media platforms' integrity.

5.2 Future Scope

Looking forward, our deepfake detection model presents several avenues for future exploration and enhancement. Firstly, further research could focus on improving the model's performance in detecting sophisticated deepfake manipulation techniques, such as those employing advanced generative adversarial networks (GANs) or leveraging novel data augmentation strategies. Additionally, integrating multi-modal information, such as audio and textual cues, could enhance the model's robustness and reliability, enabling more comprehensive analysis of multimedia content. Moreover, exploring the potential of federated learning approaches could facilitate collaborative model training across distributed datasets while preserving data privacy and security. Furthermore, efforts to enhance the model's interpretability and explainability could foster trust and transparency in the decision-making process, empowering users to better understand and contextualize the model's predictions. Lastly, ongoing advancements in hardware acceleration technologies, such as dedicated AI accelerators or edge computing platforms, present opportunities to further optimize the model's computational efficiency and scalability for real-time deployment in resource-constrained environments. By pursuing these avenues of research and innovation, our deepfake detection model can continue to evolve and adapt to emerging threats, ensuring its efficacy and relevance in combating digital deception in the years to come.

REFERENCES

[1] S. Tariq, S. Lee, and S. S. Woo, "A convolutional lstm based residual network for deepfake video detection," *arXiv preprint arXiv:2009.07480*, 2020.

[2] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.

[3] V. R. B. M. B. S. Dr. CH.V. Phani Krishna, Sowmya Arukala, "Deepfake detection using lstm and resnext," *Journal of Engineering Sciences*, pp. 1–9, 2022.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7, IEEE, 2018.

[5] D. A. Coccomini, G. K. Zilos, G. Amato, R. Caldelli, F. Falchi, S. Papadopoulos, and C. Gennaro, "Mintime: Multi-identity size-invariant video deepfake detection," *arXiv preprint arXiv:2211.10996*, 2022.

[6] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," in *2020 25th international conference on pattern recognition (ICPR)*, pp. 5012–5019, IEEE, 2021.

[7] D. A. Coccomini, N. Messina, C. Gennaro, and F. Falchi, "Combining efficientnet and vision transformers for video deepfake detection," in *International conference on image analysis and processing*, pp. 219–229, Springer, 2022.

[8] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *International journal of computer vision*, vol. 130, no. 7, pp. 1678–1734, 2022.

[9] R. Rafique, R. Gantassi, R. Amin, J. Frnda, A. Mustapha, and A. H. Alshehri, "Deep fake detection and classification using error-level analysis and deep learning," *Scientific Reports*, vol. 13, no. 1, p. 7422, 2023.

[10] A. M. Almars, "Deepfakes detection techniques using deep learning: a survey," *Journal of Computer and Communications*, vol. 9, no. 05, pp. 20–35, 2021.

[11] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, p. e2110013119, 2022.

[12] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation," in *The 20th Irish machine vision and image processing conference (IMVIP)*, pp. 133–136, 2018.

[13] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts. arxiv 2018," *arXiv preprint arXiv:1811.00656*, 1811.

[14] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann, "Media forensics considerations on deepfake detection with hand-crafted features," *Journal of Imaging*, vol. 7, no. 7, p. 108, 2021.

[15] S. Saha, R. Perera, S. Seneviratne, T. Malepathirana, S. Rasnayaka, D. Geethika, T. Sim, and S. Halgamuge, "Undercover deepfakes: Detecting fake segments in videos," *arXiv preprint arXiv:2305.06564*, 2023.

[16] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15108–15117, 2021.